UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Schuetze, B. A., & Yan, V. X. (2023). Psychology Faculty Overestimate the Magnitude of Cohen's d Effect Sizes by Half a Standard Deviation. *Collabra: Psychology*, *9*(1). https://doi.org/10.1525/collabra.74020

Methodology and Research Practice

Psychology Faculty Overestimate the Magnitude of Cohen's d Effect Sizes by Half a Standard Deviation

Brendan A. Schuetze[1][a], Veronica X. Yan[1]

[1] Department of Educational Psychology, The University of Texas at Austin, TX, US

Keywords: Statistical Cognition, Effect Size, Cohen's d, Standardized Mean Difference, Statistics Education

https://doi.org/10.1525/collabra.74020

Collabra: Psychology

Vol. 9, Issue 1, 2023

In this experiment, we recruited 261 psychology faculty to determine the extent to which they were able to visually estimate the overlap of two distributions given a Cohen's *d* effect size; and vice-versa estimate *d* given two distributions of varying overlap. In a pre-test, participants in both conditions over-estimated effect sizes by half a standard deviation on average. No significant differences in estimation accuracy by psychology sub-field were found, but having taught statistics coursework was a significant predictor of better performance. After a short training session, participants improved substantially on both tasks on the post-test, with 63% reduction in absolute error and negligible overall bias (98% bias reduction). Furthermore, post-test performance indicated that learning transferred across answering modes. Teachers of statistics might find it beneficial to include a short exercise (less than 10 minutes) requiring the visual estimation of effect sizes in statistics coursework to better train future psychology researchers.

> "And on the eighth day, Cohen said 0.20 standard deviations is small, 0.50 medium, and 0.80 large."

Psychological research—the study of human behavior, with all its multifaceted influences, interactions, and outcomes—is complex. Hence, the use of statistics is critical in finding patterns and making sense of variance. One of the most relied upon indices, for many decades, has been the infamous *p*-value, the cornerstone of "statistical significance." But *p* values, alone, give only limited insight into the results of an experiment. Rather, effect sizes are critical for contextualizing the findings of statistical tests (see Cohen, 1994), as *p* values do not indicate the magnitude or practical importance of an effect. Increasingly journal guidelines have required authors to report effect sizes to aid in interpretation. But how interpretable are these quantities? In the present study, we examine whether psychology researchers understand what effect sizes represent in terms of changes to the underlying separation between distributions. What researchers understand about effect sizes has critical implications for their understanding of the phenomena they study, how they design their studies, and for ongoing debates concerning meaningful effects in social science research.

Cohen's *d* is perhaps one of the most frequently used statistical concepts in the modern day (experimental) psychologists' toolkit after the infamous *p* value. Psychologists have been using, interpreting, and applying Cohen's *d* for decades since its formalization by Cohen (1977). Yet, an increasing emphasis on replicability in psychology over the last decade has meant that standardized effect sizes have only become more important (Szucs & Ioannidis, 2017). Durlak (2009) writes that reporting effect sizes should be seen "as an essential component of good research" (p. 918), a notion endorsed by most psychology researchers surveyed by Collins (2022).

Standardized effect sizes, such as Cohen's *d,* R^2, or odds ratios are a key component of what Cumming (2014) termed "the new statistics" that places less importance on *p* values and greater importance on interpreting confidence intervals around effects. Effect sizes are critical for psychologists to use in the power calculations that underlie the statistical planning of original studies and replication attempts (Cohen, 1977). Without a predetermined minimum effect size of interest, replication attempts are of little value, because they may be underpowered to detect the target effect at the desired power. Standardized effect sizes are also the inputs to evidence synthesis techniques, such as meta-analyses (McGrath & Meyer, 2006). Here, effects sizes are used to help compare the results of statistical tests across experiments.

Along with the increased importance placed on effect size reporting has been more considered thought given to the interpretation of standardized effect sizes. In other

a Correspondence concerning this article should be sent to Brendan Schuetze, brendan.schuetze@gmail.com

words, what constitutes a small, medium, or large effect? Historically, Cohen's (1988) guidelines have most strongly impacted research practice (see Collins & Watt, 2021a). These guidelines put forth the oft-repeated *d*s of 0.20, 0.50, and 0.80 as small, medium, and large effects, respectively. Recently, social scientists have argued that these guidelines need to be tailored to the research context, rather than adopted uncritically across the spectrum of research topics. We note that Cohen, himself, was not in favor of rigidly interpreting these guidelines (Durlak, 2009; Sawilowsky, 2009).

Thus, several attempts have been made within sub-fields to adjust the guidelines to the phenomena under study. Within education research, John Hattie's (2012) *Visible Learning* argues that only effects larger than $d = 0.40$ should be given attention by teachers, as it represents the average effect size found in his meta-synthesis on the factors of student achievement. Others, however, might argue that $d = 0.40$ is unreasonably large, as large-scale education field trials result in average effects of only $d = 0.05 - 0.17$ (Kraft, 2020). Similar points in favor of tempering expectations have been made by social scientists, such as Szaszi et al. (2022), who argue a *d* of 0.43 is "implausibly large" (p. 1) for nudge intervention; or Lovakov and Agadullina (2021) who found that the interquartile range of effect in social psychology spanned $d = 0.15 - 0.65$. Conversely, other fields might yield larger recommendations: Plonsky and Oswald (2014) found that the interquartile range of effects in second-language research spanned $d = 0.45 - 1.08$. As a note of caution, we are not necessarily endorsing the idea that "small," "medium," and "large" effects should be determined via comparison to empirical benchmarks. These benchmarks might be best decided through cost-benefit analysis in applied settings (Kraft, 2020). Just because the average large-scale educational intervention benefits students with a $d = 0.05$ does not necessarily mean this should automatically be considered a medium effect when discussing education research with policy makers.

In summary, interpreting standardized mean effect sizes is a perilous enterprise that is contextual to the field of study and the type of trial being run. However, one aspect psychologists should agree on is what different values of Cohen's *d* mean in terms of distributional overlap and separation. What does it look like when two distributions are separated by a $d = 0.50$? Are they mostly overlapping? Are they almost entirely separated? Building this sort of intuition is important for communicating and contextualizing findings to non-researcher audiences (e.g., policy makers, students, educators); it is also the focus of the present study.

What Does Cohen's *d* Represent?

Despite the increasing emphasis on effect size reporting and power analysis in psychology (Collins & Watt, 2021b), there is less known about the extent to which psychologists are familiar with and intuitively understand what a Cohen's *d* represents. That said, most researchers are probably familiar with the formula definition of Cohen's *d*:

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

In essence, a standardized mean difference, whether Cohen's *d,* Glass's Δ, or Hedge's *g*, is calculated by dividing the difference between two means—often the mean control and experimental group task performances—by a measure of the standard deviation of these distributions. We leave exact equations to the numerous statistics resources covering this information (e.g., Ferguson, 2016).

As opposed to these formal definitions of standardized mean differences, in the present study we are interested in the intuitive understanding of effect sizes that psychological researchers might have. Given their specialized training in statistics and frequent engagement with these quantities in their teaching and research, we sought to know if researchers could estimate Cohen's *d*s by simply looking at two distributions separated by an unknown *d*. Specifically, in the present study we are interested in the extent to which psychology faculty understand and visualize what a Cohen's *d* of 0.50 looks like as opposed to a $d = 0.10$ or a $d = 1.00$.

Theoretically, this task should be possible: Different Cohen's *d*s represent varying levels of overlap between the treatment and control distributions (Magnusson, 2022; Reiser & Faraggi, 1999). The overlapping coefficient can be understood as the proportion of overlap between two distributions' probability mass (Inman & Bradley, 1989). When the two distributions are normal and of equal variance, the overlapping coefficient is a transformation of Cohen's *d* (Pastore & Calcagnì, 2019), depicted in Figure 1. Under these assumptions, a $d = 0.50$ equates to 80 percent overlap, while a $d = 1.35$ equates to 50 percent overlap. Therefore, estimation of effect sizes separating two distributions should be able to be done purely perceptually—no calculations involved.

Prior research suggests that just because a statistical visualization *can be* accurately interpreted does not mean it *will be* (e.g., Kerns et al., 2020). Most existing work in statistical cognition focuses on confidence intervals and standard error interpretation. Generally, this work has found that researchers exhibit biases and certain misconceptions concerning the interpretations of statistical quantities (Cumming & Finch, 2005). For example, Cumming et al. (2004) found that researchers generally overestimated the proportion of replication means that would be captured by a 95% confidence interval (see also Belia et al., 2005). However, no studies to our knowledge have examined researcher perceptions of standardized mean differences.

Present Study

In the present study, we put this notion to the (eye) test. We recruited 261 psychology faculty from R1 universities, presented them with distributions separated by randomly varying Cohen's *d*s, and asked them to estimate the difference between the distributions. After five pre-test trials without feedback, we afforded them 15 training trials with feedback, and evaluated their performance with a final post-test of 10 additional trials. This repeated-measures design allowed us to not only estimate preconceived per-

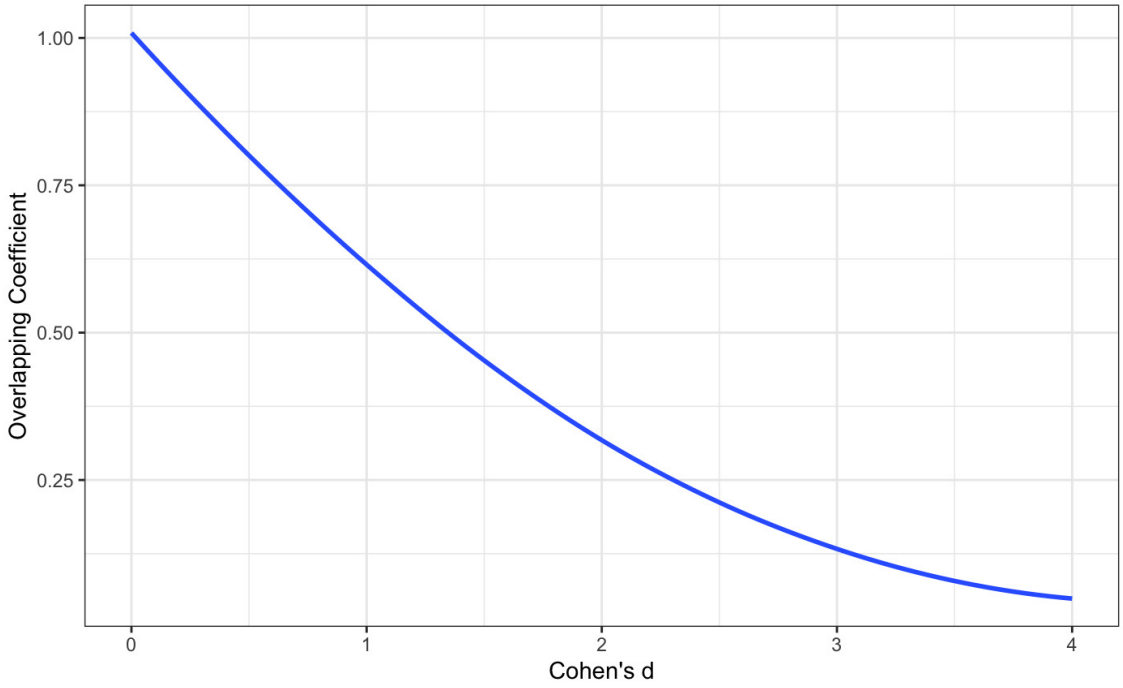

Figure 1. The Overlapping Coefficient as a Function of Cohen's d

Note. The overlapping coefficient is a statistical quantity representing the percentage of overlap between two distributions. This functional relationship depicted in the graph above assumes equal variance and normality of the distributions.

ceptions of different effect sizes, but also shows that these perceptions can be trained to greatly improve effect size estimation (reducing mean absolute error by over half and decreasing bias to essentially zero). Furthermore, inspired by Yarkoni's (2022) call for more robust experimental designs, we use two different operationalizations of the task showing largely similar results across both operationalizations. Post-test performance was assessed using two modes of answering (dragging a slider vs. typing out the *d*), showing that transfer of learning occurred across both conditions and was not merely an artifact of increased familiarity with the interface.

Methods

Participants

The study link was emailed to 4,911 faculty previously identified as teaching at a psychology or psychology adjacent department (e.g., Human Development and Family Sciences, Educational Psychology) at a Carnegie R1 "doctoral - very high research" university. These faculty were chosen due to the emphasis on research at these universities and thus the increased likelihood of needing to interpret, report, and analyze effect sizes in day-to-day research work. Furthermore, these universities tend to have doctoral programs in psychology, where many psychology PhD students are trained. Of these potential participants, 261 faculty (139 in the drag condition and 122 in the type condition) participated in the study for an overall completion rate of five percent.

For the purpose of power analysis, we note that such a sample with the repeated measures structure of the present study can detect a between-participants effect as small as $d = 0.25$ between the two conditions, given 30 repeated measures, a correlation among repeated measures of 0.50, power of .80, and alpha of .05. For within subject comparisons, such as from pre- to post-test comparisons, we had higher power to detect changes over time. An effect size as small as $d = 0.06$ could be detected given three time points, 30 repeated measures, a correlation among repeated measures of 0.50, power of .80, and alpha of .05. These power analyses assume a repeated-measures ANOVA and are presented for informational purposes. They were not used to guide the recruitment of the present study, as the response rate to the experiment was not within the experimenters' control.

The average year of PhD (or highest degree) completion was 2002 (*Mdn* = 2005), with a range between 1966 and 2022. 156 participants were men, 95 women, and 10 participants' gender was non-binary, preferred not to state, or N/A. All participants were recruited in the late summer and early fall of 2022 via an initial recruitment email and a reminder email if they did not complete the study after approximately one week. See Table 1 for full sub-field and other demographics. This study was approved and overseen by the Institutional Review Board of the authors' university.

Stimuli

Participants were shown two distributions defined by varying standard deviations and separated by an effect size between $d = 0.02$ (equivalent to $R^2 < .01$) and 2.00 ($R^2 = 0.50$). This range of effects was chosen to more than cover the range of commonly-observed effect sizes in psychology

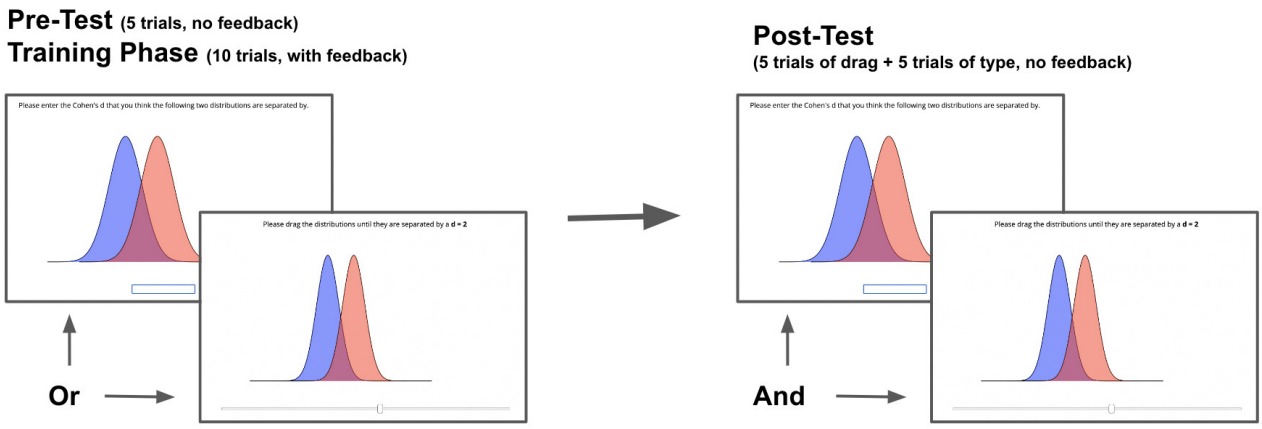

Figure 2. Experiment Summary

Note. The method of answering (drag distributions vs. type Cohen's *d*) was manipulated between participants and stayed constant throughout the pre-test and training portions of the experiment session. Conversely, all participants answered five trials using each answering method in the post-test, regardless of condition assignment to assess transfer.

studies. Funder and Ozer (2019) indicate that an $R^2 > 0.16$ is "a very large effect size (r = .40 or greater) in the context of psychological research is, we suggest, likely to be a gross overestimate that will rarely be found in a large sample or in a replication" (p. 166). For this reason, we did not ask participants to estimate Cohen's *d* effect sizes greater than $d = 2$, as this is already an unlikely large effect in the type of research performed by experimental psychologists. We excluded Cohen's *d*s exactly equal to 0 from our stimuli set, as they would be defined by a complete overlap of the two focal distributions, which may have confused participants. The different magnitudes of effect sizes shown to participants in each trial were randomly drawn from a uniform distribution with the constraint that no exact effect size could be repeated twice per participant. Thus, each participant saw an entirely distinct range and order of effect sizes across the 30 total trials they completed. We also randomly manipulated the standard deviation (aspect ratio) of the stimuli from trial-to-trial using one of four pre-sets varying in distribution width. This was particularly important in the drag condition, so that locations on the slider itself did not map directly onto a specific Cohen's *d* (see Figure 2). Therefore, participants could not simply, for example, memorize that the middle of the slider was equal to a *d* of 0.50.

Procedure

Once informed consent was gathered from participants via Qualtrics survey, participants were redirected to a webpage programmed in jsPsych (de Leeuw, 2015). The experiment portion of the study was composed of two conditions (type vs. drag), with condition membership randomly assigned between participants. In the type condition, participants were presented with two distributions separated by an effect size drawn from a uniform distribution ranging between $d = 0.02$ and 2.00, and asked to type out the effect size in numeric form. In other words, the type condition required participants to judge the effect being shown to them and type it out in a textbox (i.e., "Please enter Cohen's *d* that you think the following two distributions are separated by."). In the drag condition, participants were given the opposite task. They were told an effect size in numeric form and asked to use a slider to drag the two distributions apart such that they correctly represented the numeric effect size (e.g., "Please drag the distributions until they are separated by a $d = 0.5$"). All trials were self-paced, only advancing after an answer was given and the next button was pressed. There were three phases to the experiment: Pre-Test, Training, and Post-Tests.

Pre-Test

The pre-test phase started the experiment with five trials without feedback to estimate the participants' pre-existing ability to judge effect sizes of varying magnitudes. After the training phase concluded, participants made a Likert judgment relating the confidence in their ability to estimate the effect sizes ("I feel confident in my ability to estimate and visualize different Cohen's *d* effect sizes"). This judgment was made on a five-point scale from strongly disagree to strongly agree.

Training Phase

The training phase task was the same as that of the pre-test phase, except for the inclusion of immediate feedback after each trial. The training phase lasted 15 trials total. Although the answer submission phase of the trials was self-paced, feedback was not. Feedback was presented regardless of correctness and always lasted five seconds. On the feedback screen, participants were told their response in Cohen's *d* units, the correct response in Cohen's *d* units, and the error between these two responses. They were also shown the distribution stimuli from the relevant trial. After the 15 trials elapsed and this phase concluded, participants were once again asked for their confidence using the same five-point Likert question as previously described.

Post-Test

Then, participants entered the post-test phase, where they completed the same task as before without feedback to get one last estimate of their ability to judge effect sizes. However, this time participants were also exposed to five opposite condition-type trials of which they were assigned. In other words, for the final phase all participants completed five type trials and five drag trials, regardless of condition assignment. By testing all participants under both formats, we could assess transfer and ensure that final test results were more than just interface familiarity effects. The order of the type versus drag trials was randomly shuffled between participants.

Demographic Collection

After these last ten trials, participants complete demographic questions. Specifically, they indicated the primary type of effect size used in their research (e.g., Cohen's *d,* odds ratios, variance explained measures), time in years since earning their PhD, their primary subfield of psychology (e.g., cognitive, social or personality, quantitative; see Table 1), and whether they had ever taught statistics classes. They also indicated their gender and age. Upon completing the demographic portion of the study, participants were thanked for their time and the study completed.

Results

Pre-registration and Data Cleaning

The following analyses were pre-registered unless otherwise specified. Pre-registration documentation and open data, except for age and gender information which has been removed to ensure anonymity of participants, is hosted on OSF at: https://osf.io/jxw8t/. Only datasets with complete experimental data were used for analytic purposes (participants were allowed to leave demographic questions blank and this was not grounds for removal). One trial was removed from the analyses because of the pre-registered criterion that any response greater than |6| z-scores from the mean of an individual's responses would be removed to avoid the undue influence of typos (e.g., d = 200 instead of 2.00). In a deviation from our pre-registration, one participant was removed from the dataset prior to our analyses, because they answered with several Cohen's *d*s between 200 and 500. This was the sole participant to type out answers above d > 10. Because this participant made this error several times (dragging their individual mean up), the z-score rule did not filter their answers. This participant was excluded from demographics and participant information reported earlier in the paper.

Average Bias and Absolute Error Before Training

Both average bias (error; Figures 3 and 4) and absolute error (|bias|; Figure 5) changed over the course of the experiment. Before receiving feedback, participants in both the drag and type conditions over-estimated the gap between distributions for a given effect size. The naive estimate of average bias in the pre-test phase was equal to 0.02 (i.e., essentially zero). However, this estimate does not account for the fact that the two conditions are inverse tasks, and show equal, but opposite biases (of approximately 0.5SD). In the pre-test, bias was fairly consistent across all ranges of the correct answer for the drag condition, whereas in the type condition bias was higher at the upper-end of the distribution of correct answers (see Figure 3).

After reverse-scoring the dependent variable (error) in the type condition to make the conditions' bias comparable, we estimated an intercept-only model accounting for participants as a random intercept (see OSF for full regression model syntax and output). Pre-test bias was significantly different from zero, b = 0.51, SE = 0.04, p < .001, and varied from participant to participant in magnitude (τ^2_{ID} = 0.33, ICC = 0.59). Altogether, 80% of participants exhibited an upward bias on the pre-test (M_{bias} > 0).

In other words, when presented with two distributions, participants tended to think these distributions were separated by a Cohen's d that was half a standard deviation smaller than reality. And conversely, when asked to drag the distributions apart, they dragged the distributions one-half standard deviation too far. This indicates that regardless of answer modality (type or drag condition) participants over-estimated how large Cohen's d effect sizes were in terms of the true separation of distributions. Analysis showed that the type condition showed slightly lower absolute error than the drag condition in the pre-test, b = -0.12, SE = 0.06, p = .03 (see Figure 5).

Average Bias and Absolute Error Post-Training

As seen in Figures 4 and 5, both bias and mean absolute error were greatly reduced within five trials of training with feedback. Mean bias was nearly zero across both conditions in the 10 post-test trials (Type M_{Bias} = -0.03 standard deviations; Drag M_{Bias} = -0.04). Mean absolute error stabilized at approximately a quarter of a standard deviation for both conditions (Type M_{MAE} = 0.25; Drag M_{MAE} = 0.26).

The reduction in error and bias from pre to post-test was significant for both bias and absolute error across both conditions (ps < .001). Mean absolute error was reduced by approximately 63% and bias was reduced by approximately 98%. There was no statistically significant difference in overall accuracy between conditions (type vs. drag) in the training and post-test phase (ps > .05). This lack of difference in the post-test phase indicates that neither training condition transferred significantly better than the other (see Figure 6 and Figure 7).

Differences in Accuracy by Sub-Field of Psychology and Demographic Characteristics

We were interested in determining the extent to which differences in sub-field of psychology predicted differences in ability to estimate different Cohen's d effect sizes as measured by the pre-test prior to receiving feedback. To do this we created a hierarchical linear model predicting mean absolute error using condition, standard deviation of the stimulus (centered), year of PhD (centered), research

Table 1. Participant Demographics

	Drag (*n* = 139)	Type (*n* = 122)	Overall (*N* = 261)
Gender			
Female	50 (36.0%)	45 (36.9%)	95 (36.4%)
Male	85 (61.2%)	71 (58.2%)	156 (59.8%)
Other or NA	4 (2.9%)	6 (4.9%)	10 (3.8%)
Research Area			
Applied	12 (8.6%)	13 (10.7%)	25 (9.6%)
Clinical & Counseling	21 (15.1%)	26 (21.3%)	47 (18.0%)
Cognitive	36 (25.9%)	23 (18.9%)	59 (22.6%)
Developmental	11 (7.9%)	14 (11.5%)	25 (9.6%)
Neuroscience & Biopsychology	14 (10.1%)	16 (13.1%)	30 (11.5%)
Quantitative	9 (6.5%)	4 (3.3%)	13 (5.0%)
Social & Personality	27 (19.4%)	21 (17.2%)	48 (18.4%)
Other or NA	9 (6.5%)	5 (4.1%)	14 (5.4%)
Taught Statistics			
Yes	78 (56.1%)	64 (52.5%)	142 (54.4%)
No	61 (43.9%)	56 (45.9%)	117 (44.8%)
Other or NA	0 (0%)	2 (1.6%)	2 (0.8%)
Year of PhD			
Mean (SD)	2001 (12.51)	2002 (11.83)	2002 (12.19)
Median [Min, Max]	2004 [1969, 2019]	2005 [1966, 2022]	2005 [1966, 2022]
Missing	1 (0.7%)	5 (4.1%)	6 (2.3%)
Age			
Mean (SD)	50.4 (12.3)	48.3 (11.5)	49.4 (11.9)
Median [Min, Max]	47.0 [31, 83]	46.0 [28, 82]	46.0 [28, 83]
Missing	2 (1.4%)	10 (8.2%)	12 (4.6%)

area (dummy-coded), and whether participants had taught statistics classes before (also dummy coded) as predictors. Standard deviation (aspect ratio) of the stimulus was also incorporated as a random slope, with each participant as an intercept. Overall, we found that year of PhD was a significant predictor of mean absolute error (b = 0.005, SE = 0.002, p = .03), indicating that more recent PhD graduation was associated with very slightly worse performance as assessed by mean absolute error. We also found that teaching statistics was a significant predictor of reduced mean absolute error (b = -0.12, SE = .05, p = .02).

To test the omnibus hypothesis of the predictive importance of the research area, we calculated a reduced model, which was the same as the full model, except that it did not include the dummy-coded research area predictors (although we pre-registered the full model, we did not pre-register the exact omnibus test of interest). Then we computed a likelihood ratio test between the full and reduced models, finding no significant increase in likelihood from adding research area, $\chi^2(7) = 6.95$, p = .43. Overall, we found that researchers self-identifying as quantitative psychologists performed the best (least mean absolute error and lowest bias) on the pre-test. Due to limited sub-group sample sizes, we do not report pairwise post-hoc tests, but show descriptive statistics and confidence intervals in Table 2.

Self-Rated Confidence Over the Course of the Experiment

Three confidence judgments were made: once after the pre-test, after the training phase, and after the post-test on a 1-5 scale. Participants reported moderate confidence after the pre-test (M = 2.73, Mdn = 3), which only increased slightly after the training (M = 3.07, Mdn = 3), and post-test phases concluded (M = 2.98, Mdn = 3).

General Discussion

There is good news and there is bad news. As for the bad: in pre-training our expert researcher-participants greatly overestimated the separation of distributions for any given Cohen's d. On average, they were off by over half a standard deviation, regardless of question format. For example, when asked to separate two distributions by a Cohen's d of 0.50, they dragged the distributions to a Cohen's d of 1.00; when presented with two distributions representing a Cohen's d of 0.50, they typed their estimate as being closer to zero (see Figure 3). The implication is that researchers often think that their effect sizes reflect larger differences between distributions than has actually been found. It may be that common "rules of thumb" language that categorizes effects as small, medium, and large, mis-

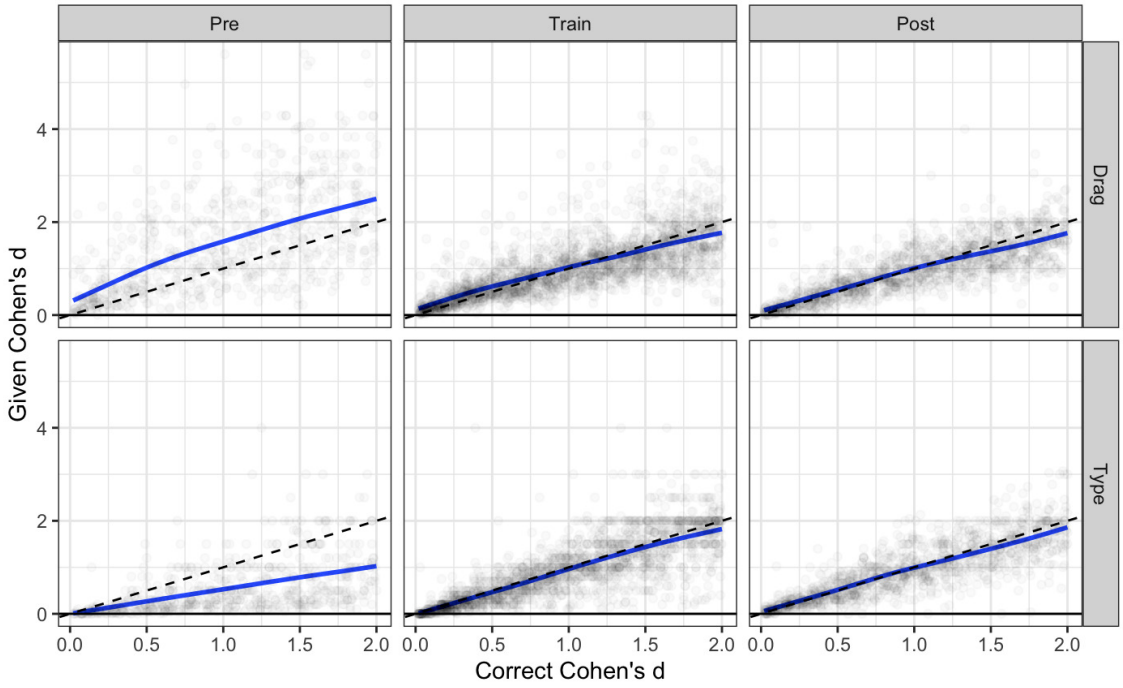

Figure 3. Given Answer by Correct Answer in Each Experiment Phase

Note. The dashed black line shows no bias (slope = 1, intercept = 0). The blue line shows the empirical line of best fit as estimated by LOESS for each condition (drag vs. type) by trial phase (pre-test, train, post-test). Thus, the difference between the two lines indicates bias.

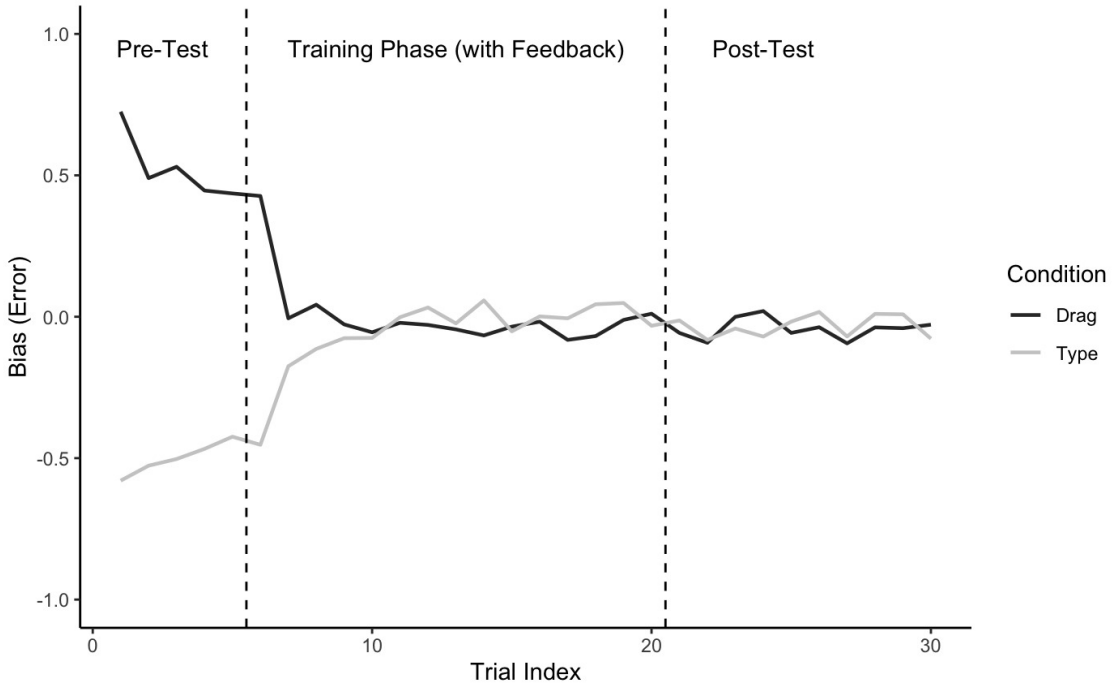

Figure 4. Mean Bias (Error) Across the Experiment

leads both researchers and the broader public, shaping understanding of these effect sizes in a way that is detached from the statistics.

This discrepancy means potential mistranslations of research to the broader audience, making claims about "profound" effects when they may not be all that profound at all. For example, a researcher excited about an effect size of $d = 0.30$ might overstate how much of a difference this represents to a policymaker. The policymaker might become disillusioned with psychological interventions when they invest potentially millions of dollars (Sims, 2020) implementing it and find a muted effect. None of this means that we should not report effect sizes; rather, that we should be

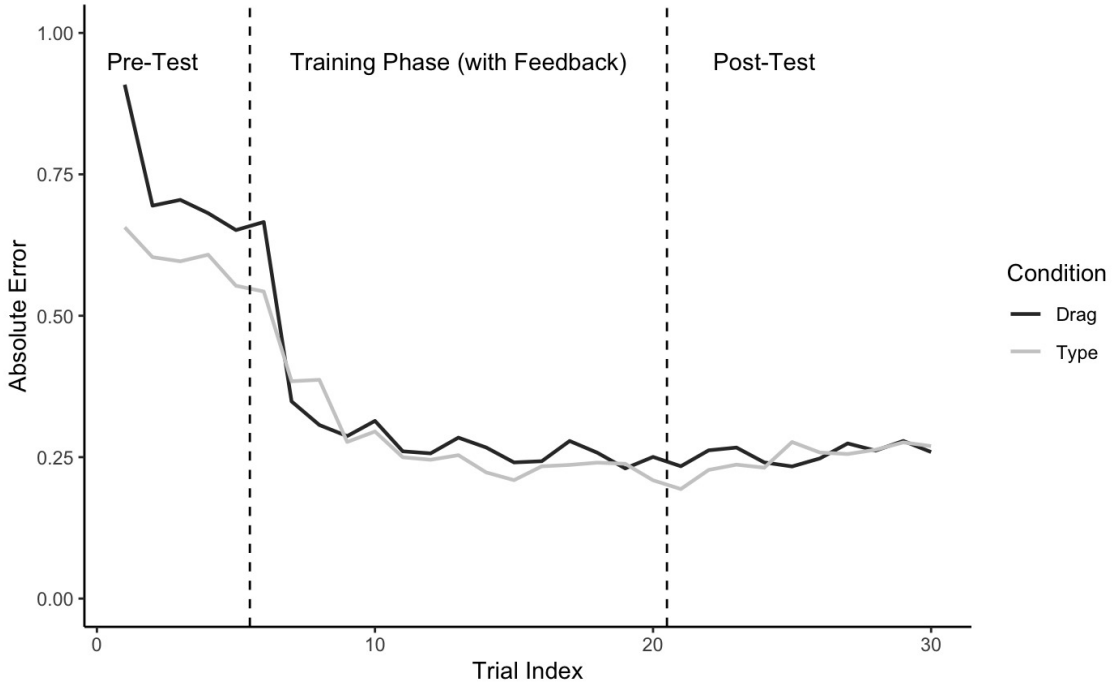

Figure 5. Mean Absolute Error (|Bias|) Across the Experiment

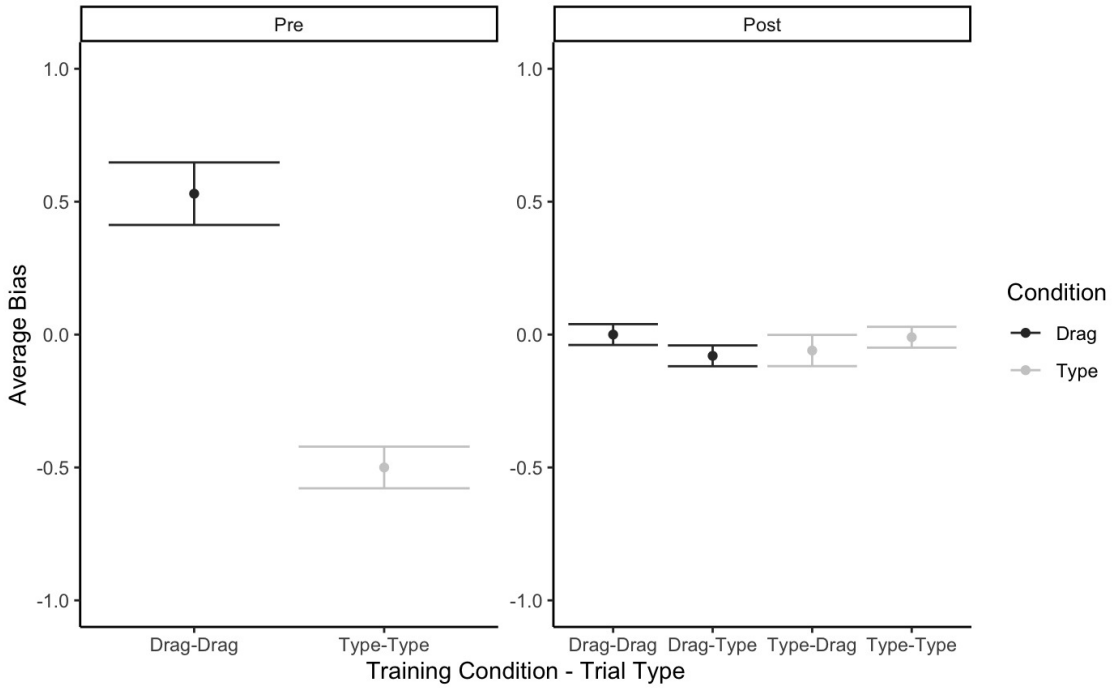

Figure 6. Mean Bias (Error) by Condition and Pre- or Post-Test

Note. Given that the type and drag conditions are inverse of one another, it is expected that they show opposite biases in terms of the participants' given Cohen's *d*. Error bars represent 95% confidence intervals after averaging within individuals.

more careful in the ways we communicate the understanding of them.

As for good news, we found that this overestimation bias was reduced to near-zero through a short intervention built on cognitive science of learning principles (e.g., many varied practice trials with feedback; see perceptual learning literature, Kellman & Massey, 2013). Mean absolute error was reduced to about 0.25 standard deviations in both conditions (see Figure 8 for the visual representation), which may represent a functional limit to this benefit of this type of training. Furthermore, we found that transfer occurred in both directions—from the type condition to the drag condition and vice-versa—indicating that the exact format was

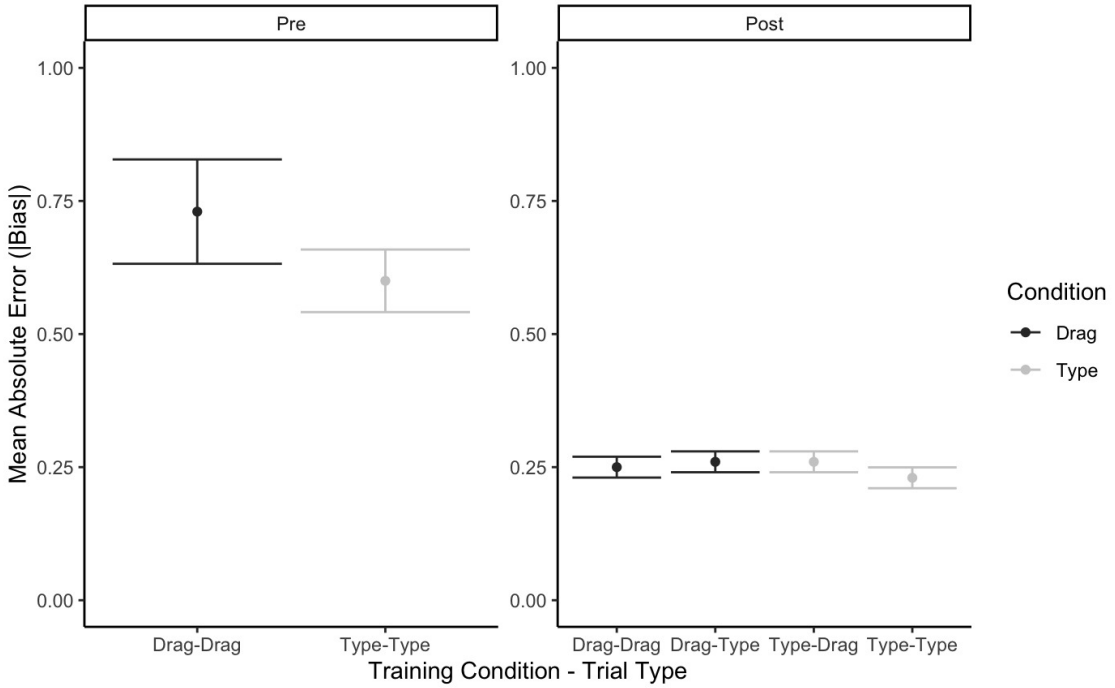

Figure 7. Mean Absolute Error by Condition and Pre- or Post-Test

Note. Knowledge transferred across conditions, with absolute error essentially constant across all types of tests (drag or type) regardless of initial training condition (drag or type). Error bars represent 95% confidence intervals after averaging within individuals.

Table 2. Pre-Test Absolute Error by Research Area

Research Area	N	Mean MAE	MAE SE	MAE 95% Confidence Interval		Mean Bias	Bias SE	Bias 95% Confidence Interval	
				Lower bound	Upper bound			Lower bound	Upper bound
Quantitative	13	0.38	0.07	0.25	0.51	0.24	0.10	0.05	0.43
Applied	25	0.56	0.06	0.43	0.68	0.42	0.09	0.23	0.60
Social & Personality	48	0.65	0.06	0.52	0.77	0.51	0.08	0.35	0.68
Cognitive	59	0.66	0.06	0.54	0.77	0.49	0.08	0.33	0.65
Neuroscience & Biopsychology	30	0.70	0.10	0.51	0.90	0.44	0.14	0.17	0.71
Clinical & Counseling	47	0.72	0.05	0.62	0.83	0.60	0.08	0.44	0.76
Developmental	25	0.79	0.11	0.58	1.00	0.63	0.14	0.35	0.91
Other or N/A	14	0.81	0.18	0.46	1.15	0.69	0.21	0.29	1.10

Note. Lower absolute error (MAE) indicates better performance. Bias closer to zero indicates better performance (positive bias means overestimating the amount of separation associated with Cohen's *d*s on average). Applied psychology includes researchers identifying as Educational, Human Factors, and I-O psychologists. Confidence intervals are +/- 1.96 SEs and are not corrected for multiple comparisons.

largely irrelevant to creating transferable knowledge of Cohen's *d* effect sizes.

That said, one limitation of the present study is that we did not assess long-term learning. Generally similar interventions have shown relatively stable learning over the long-term (Kellman et al., 2010). Moreover, our results provide evidence that training (broadly construed) can last: In the pre-test, quantitative psychologists and those who have taught statistics courses showed smaller-than-average errors (though we did not find a significant omnibus test of research area).

Finally, we found that researchers' metacognitive confidence did not increase commensurately with their performance over the course of the task. Confidence was low throughout the task. However, we did not measure confidence before the pre-test, so we do not know if confidence would have been higher before people engaged with the task. On one hand, the combination of low confidence and low initial accuracy means that researchers are reporting effect sizes without a solid understanding of what they are reporting. On the other hand, low confidence errors are better than high confidence errors. The persisting low confidence post-training might be due to the lack of didac-

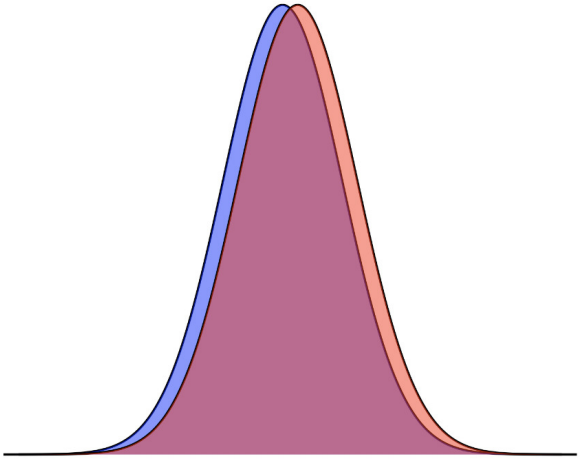

Figure 8. An Illustration of a *d* = 0.25 Effect Size

tic training within the study. Future research might benefit from combining the present intervention with a more didactic approach and further intuition building exercises aimed at helping people understand different effect sizes (e.g., odds ratios, R^2) or Cohen's *d* effect sizes under more realistic scenarios (e.g., skewed or distributions with unequal variances).

To help researchers better understand Cohen's *d* effect sizes as they relate to the overlap between distributions, we have developed an R Shiny application, which can be found at this paper's OSF link: https://osf.io/jxw8t/. This app allows for learners to practice guessing and checking their estimates of standardized mean effect sizes in a similar manner to that carried out in the present study. We also suggest readers engage with Magnusson's (2022) web app as a useful intuition-builder for better understanding standardized mean effect sizes.

Ultimately, we reiterate that effect sizes are an important aspect of statistical training for researchers in the social sciences, and it is critical that researchers know what these quantities represent, not just mathematically but in a more intuitive sense. As one of our participants wrote, "The truth is: No one ever told me that I needed this skill." Fortunately, it is not a hard skill to learn, and it seems that learning transfers fairly easily across answering modalities.

Author Contributions

Contributed to conception and design: BAS, VXY
Contributed to acquisition of data: BAS, VXY
Contributed to analysis and interpretation of data: BAS
Drafted and/or revised the article: BAS, VXY
Approved the submitted version for publication: BAS, VXY

Competing Interests

There are no conflicts of interest.

Data Accessibility Statement

Pre-registration, an R Shiny app, and open data for this study can be found at the following Open Science Framework project: https://osf.io/jxw8t/

Submitted: February 10, 2023 PDT, Accepted: March 27, 2023 PDT

References

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*(4), 389–396. https://doi.org/10.1037/1082-989x.10.4.389

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press.

Cohen, J. (1994). The earth is round (p <.05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066x.49.12.997

Collins, E. (2022). *The ideal psychologist vs. A messy reality: Using and misunderstanding effect sizes, confidence intervals and power* [Doctoral dissertation, University of Stirling]. http://hdl.handle.net/1893/34366

Collins, E., & Watt, R. (2021a). Using and understanding power in psychological research: A survey study. *Collabra: Psychology*, *7*(1), 28250. https://doi.org/10.1525/collabra.28250

Collins, E., & Watt, R. (2021b). Use, knowledge and misconceptions of effect sizes in psychology. *PsyArXiv*. Preprint. https://doi.org/10.31234/osf.io/r7vmf

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170–180. https://doi.org/10.1037/0003-066x.60.2.170

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, *3*(4), 299–311. https://doi.org/10.1207/s15328031us0304_5

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917–928. https://doi.org/10.1093/jpepsy/jsp004

Ferguson, C. J. (2016). An effect size primer: A guide for clinicians and researchers. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (4th ed., pp. 301–310). American Psychological Association. https://doi.org/10.1037/14805-020

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Inman, H. F., & Bradley, E. L., Jr. (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics - Theory and Methods*, *18*(10), 3851–3874. https://doi.org/10.1080/03610928908830127

Kellman, P. J., & Massey, C. M. (2013). Perceptual learning, cognition, and expertise. In *Psychology of Learning and Motivation* (Vol. 58, pp. 117–165). Elsevier. https://doi.org/10.1016/b978-0-12-407237-4.00004-9

Kellman, P. J., Massey, C. M., & Son, J. Y. (2010). Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. *Topics in Cognitive Science*, *2*(2), 285–305. https://doi.org/10.1111/j.1756-8765.2009.01053.x

Kerns, S., Kim, H., Grinspoon, E., Germine, L., & Wilmer, J. (2020). Toward a science of effect size perception: The case of introductory psychology textbooks [Conference presentation abstract]. *Journal of Vision*, *20*(11), 1185. https://doi.org/10.1167/jov.20.11.1185

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189x20912798

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

Magnusson, K. (2022). *Interpreting Cohen's d effect size: An interactive visualization* [Web App]. https://rpsychologist.com/cohend/

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods*, *11*(4), 386–401. https://doi.org/10.1037/1082-989x.11.4.386

Pastore, M., & Calcagnì, A. (2019). Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in Psychology*, *10*, 1089. https://doi.org/10.3389/fpsyg.2019.01089

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Reiser, B., & Faraggi, D. (1999). Confidence intervals for the overlapping coefficient: The normal equal variance case. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *48*(3), 413–418. https://doi.org/10.1111/1467-9884.00199

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 597–599. https://doi.org/10.22237/jmasm/1257035100

Sims, S. (2020). Informing better trial design: A technical comment on Lortie-Forgues and Inglis (2019). *Educational Researcher*, *49*(4), 289–290. https://doi.org/10.3102/0013189x19867931

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S., & Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences*, *119*(31), e2200732119. https://doi.org/10.1073/pnas.2200732119

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.3001151

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/s0140525x20001685

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/74020-psychology-faculty-overestimate-the-magnitude-of-cohen-s-d-effect-sizes-by-half-a-standard-deviation/attachment/155084.docx?auth_token=Taz5x35qan64GYH3yXQp